Journal of Artificial Intelligence and Engineering Applications
Websites:https://ioinformatic.org/

15th October 2021. Vol. 1. No. 1

Analysis of Decreased Public Awareness in the Application of Health Protocols with the C4.5 . Algorithm

Retno Arfika[1]*, M.Safii[2]

[1]STIKOM Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia*
[2]AMIK Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia
[3]Jln Sudirman Blok A No.1-3 Pematangsiantar, North Sumatra
** retnoarfika03@gmail.com, * m.safii@amiktunasbangsa.ac.id*

Abstract

The purpose of this study is to determine the dominant factors that affect the decline in public awarenesson the application of health protocols using the C4.5 Algorithm. Sources of data used in this study obtained by conducting observations and interviews. The variables used include (1) Employment, (2) Environment, (3) Sanctions and (4) Concern. The research test process uses RapidMiner software to create a decision tree. The results obtained 6 rules with 4 rules decreasing status and 2 rules increasing status. The level of accuracy obtained is 100%. The results of this study are expected to be input for the surrounding community to better understand the importance of implementing health protocols at this time, so that they can help the Government to succeed in the health protocol awareness program in inhibiting the spread of Covid-19 in Indonesia.

Keywords: *C4.5 Algorithm, Covid-19, Data Mining, Society*

1. Introduction

The implementation of health protocols is an important thing that must be done in preventing or protecting themselves from the spread of the Corona virus, but so far most people think this is not an important thing that must be done continuously. There are several possibilities that influence the community to no longer be disciplined in applying health protocols, such as in terms of work, the environment, the application of sanctions in the community and public concern in responding to the Covid-19 cases that occur. During the determination of Covid-19 as a pandemic in Indonesia, many people experienced fear, their health declined, and an excessive sense of worry made their immune system weak. Below are various government policies to prevent the spread of the corona virus. Social distancing policy, physical distancing policy, personal protective equipment (mask) policy, personal hygiene policy (hand washing), work/study from home policy, postponing policy all activities that gather large crowds, the Large-Scale Social Restriction Policy (PSBB); until the last, New Normal policy implementation[1]. Health protocols that are enforced in almost every place are rules that need to be obeyed by the community so that people can carry out their activities safely. Adhering to this protocol does not completely protect the public from the Corona virus but can prevent exposure to the virus itself. The high number of transmission of Covid-19 cases in Indonesia is still caused by various problems. These problems are especially common among the community. One of them is the lack of concern, awareness, and knowledge of the public about the threat of the Covid-19 pandemic[9].

However, people are starting to ignore health protocols. For example, roaming outside without wearing a mask, crowding around without thinking about the importance of keeping a distance, shopping at supermarkets or eating outside the home and ignoring the health protocols that have been provided on the spot and some accidentally don't wear masks for some reason. The 3M health protocol aims to the community can still carry out their activities safely, without endangering the health and safety of themselves and others. In its implementation, the implementation of the 3M health protocol has various responses from residents, many of whom comply with it with full awareness, but some do not care. It is the people's disobedience that exacerbates the situation[10]. In this study, one of the Data Mining techniques is used, namely the C4.5 classification algorithm. C4.5 algorithm. is a group of decision tree algorithms (decision tree). This algorithm has input in the form of training samples and samples. Training samples in the form of sample data that will be used to build a tree that has been tested for truth[2]. While the decision tree is the result of the C4.5 algorithm where the decision tree is useful for exploring data, finding hidden relationships between a number of candidate input variables and a target variable.[3]. The decision tree is able to make complex decisions simpler, so that decision makers will better interpret solutions to problems [8]. Previous researchers who have conducted a study using the C4.5 Algorithm with the title C4.5 Classification Analysis of the Factors Causing the Decline of Student Learning Achievement in the Pandemic Period. This study was conducted to determine the obstacles in the learning process that affect student achievement.[4].

2. Results and Discussion

2.1. Data Mining

Data mining is a process that employs one or more computer learning techniques (machine learning) to analyze and extract knowledge automatically.[5]. The C4.5 algorithm is one of the methods in data mining. In general, the steps of the C4.5 algorithm to create a decision tree is to start by selecting the attribute as the root (rootnode).
basic formula for entropy.`

$Entrophy\ (S) = -\sum pi\ log2\ (pi)$ (1)
Where :
S : case set
n : number of partitions S
pi : the proportion of Si to S
then calculate the gain for object separation.
By using the following formula:
$(S, A) = Entropi\ (S) - \sum |si|\ |s|\ n\ i{=}1\ X\ Entropi\ (Si)$ (2)
Where :
S : data set
A : attribute
n : number of attribute partition A
|Si| : number of cases on partition i
|S| : number of cases in S
Then create a branch and split the cases in branches

2.2. Decision Tree

The decision tree is the result of the Entropy and information gain calculation process, after repeated calculations until all tree attributes have classes and the calculation process can no longer be carried out[2].

2.3. Rapid Miner

RapidMiner is an open source application that can be used to perform data mining processes[6]. Rapid Miner is a science data software platform developed by the company of the same name that provides an integrated environment for machine learning, deep learning, text mining, and predictive analytics.[7].

2.4. Research data

In this study, the data used were the results of the distribution of questionnaires given to the community as many as 100.
In determining the classification of the class value on each attribute can be seen in the table below.

Table 1: Data Class Type

Attribute	Field Name	Data Class
C1	Work	Disturbed - Not Disturbed
C2	Environment	Support Not Support
C3	Penalty	Firm - Less Firm
C4	Concern	Care - Careless - Don't not mather

The following is a sample of the data resulting from the questionnaire recapitulation processed in Microsoft Excel in Table 3.1.

Table 2: Questionnaire Recapitulation Data

No	**PROFESSION**	**ENVIRONMENT**	**PENALTY**	**CONCERN**	**Awareness**
R1	Disturbed	Does not support	Lack of firm	Does not matter	Decrease
R2	Disturbed	Does not support	Firm	Careless	Decrease
R3	Not disturbed	Support	Firm	Care	Increase
R4	Disturbed	Does not support	Lack of firm	Careless	Decrease
R5	Not disturbed	Support	Firm	Careless	Decrease
R6	Disturbed	Does not support	Lack of firm	Careless	Decrease
R7	Disturbed	Does not support	Lack of firm	Does not matter	Decrease
R8	Disturbed	Does not support	Lack of firm	Careless	Decrease
R9	Disturbed	Support	Firm	Careless	Decrease
R10	Disturbed	Support	Lack of firm	Does not matter	Decrease
R11	Disturbed	Does not support	Lack of firm	Careless	Decrease
R12	Disturbed	Support	Firm	Careless	Decrease
R13	Disturbed	Support	Firm	Does not matter	Decrease
R14	Disturbed	Does not support	Firm	Care	Decrease
R15	Disturbed	Does not support	Lack of firm	Careless	Decrease
R16	Not disturbed	Does not support	Lack of firm	Careless	Decrease
R17	Disturbed	Does not support	Lack of firm	Does not matter	Decrease
R18	Disturbed	Does not support	Lack of firm	Careless	Decrease
R19	Disturbed	Does not support	Lack of firm	Does not matter	Decrease
R20	Disturbed	Support	Lack of firm	Careless	Decrease
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
R100	Disturbed	Does not support	Lack of firm	Careless	Decrease

.

2.5. C4.5 Algorithm Calculation

Calculating the total entropy:

Entropy [Total] $= \left(-\frac{87}{100}\ x\ log_2\left(\frac{87}{100}\right)\right) + \left(-\frac{13}{100} x log_2\left(\frac{13}{100}\right)\right)$
$= 0.557438185$

Calculating entropy and gain (C1) :

Entropy [Disturbed] $= \left(-\frac{69}{72}\ x\ log_2\left(\frac{69}{72}\right)\right) + \left(-\frac{3}{72} x log_2\left(\frac{3}{72}\right)\right)$
$= 0.249882293$

Entropy [Not disturbed] $= \left(-\frac{18}{28}\ x\ log_2\left(\frac{18}{28}\right)\right) + \left(-\frac{10}{28} x log_2\left(\frac{10}{28}\right)\right)$
$= 0.940285959$

Gain $= 0.557438185 - \left(\left(\frac{72}{100} x\ 0{,}249882293\right) + \left(\frac{28}{100} x\ 0{,}940285959\right)\right)$
$= 0.114242866$

Calculating entropy and gain (C2) :

Entropy [Support] $= \left(-\frac{14}{25}\ x\ log_2\left(\frac{14}{25}\right)\right) + \left(-\frac{11}{25} x log_2\left(\frac{11}{25}\right)\right)$
$= 0.989587521$

Entropy [No Support] $= \left(-\frac{73}{75}\ x\ log_2\left(\frac{73}{75}\right)\right) + \left(-\frac{2}{75} x log_2\left(\frac{2}{75}\right)\right)$
$= 0.177389453$

Gain $= 0.557438185 - \left(\left(\frac{25}{100} x\ 0{,}989587521\right) + \left(\frac{75}{100} x\ 0{,}177389453\right)\right)$
$= 0.176999215$

Calculating entropy and gain (C3) :

Entropy [Firm] $= \left(-\frac{13}{24}\ x\ log_2\left(\frac{13}{24}\right)\right) + \left(-\frac{11}{24} x log_2\left(\frac{11}{24}\right)\right)$
$= 0.994984828$

Entropy [Lack of firm] $= \left(-\frac{74}{76}\ x\ log_2\left(\frac{74}{76}\right)\right) + \left(-\frac{2}{76} x log_2\left(\frac{2}{76}\right)\right)$
$= 0.175565026$

Gain $= 0.557438185\ \left(\left(\frac{24}{100} x\ 0{,}994984828\right) + \left(\frac{76}{100} x\ 0{,}175565026\right)\right)$
$= 0.185212407$

Calculating entropy and gain (C4) :

Entropy [Care] $= \left(-\frac{7}{20}\ x\ log_2\left(\frac{7}{20}\right)\right) + \left(-\frac{13}{20} x log_2\left(\frac{13}{20}\right)\right)$
$= 0.934068055$

Entropy [Careless] $= \left(-\frac{64}{64}\ x\ log_2\left(\frac{64}{64}\right)\right) + \left(-\frac{0}{64} x log_2\left(\frac{0}{64}\right)\right)$
$= 0$

Entropy [Does not matter] $= \left(-\frac{16}{16}\ x\ log_2\left(\frac{16}{16}\right)\right) + \left(-\frac{0}{16} x log_2\left(\frac{0}{16}\right)\right)$
$= 0$

Gain $= 0.557438185 - \left(\left(\frac{20}{100} x\ 0{,}934068055\right) + \left(\frac{64}{100} x\ 0\right) + \left(\frac{16}{100}\ x\ 0\right)\right)$
$= 0.3706246$

Attribute C4 gets the highest gain so that it is used as a node. Because the Peduli attribute class has not yet obtained the results, further calculations must be carried out. Meanwhile, the Less Caring and Not Caring attribute classes have obtained the same result, namely Decreasing. The following are the results of the calculation of the entropy and gain values described in Table 3.

Table 3: Node Calculation Results 1.

Node 1		**Amount**	**Decrease**	**Increase**	**Entrophy**	**Gain**
		100	87	13	0.557438185	
C1						0.114242866
	Disturbed	72	69	3	0.249882293	
	Not disturbed	28	18	10	0.940285959	
C2						0.176999215
	Support	25	14	11	0.989587521	
	Does not support	75	73	2	0.177389453	
C3						0.185212407
	Firm	24	13	11	0.994984828	
	Lack of firm	76	74	2	0.175565026	
C4						**0.3706246**
	Care	20	7	13	0.934068055	
	Careless	64	64	0	0	
	Does not matter	16	16	0	0	

Table 4: Node Calculation Results 1.1

Node 1.1		**Amount**	**Decrease**	**Increase**	**Entrophy**	**Gain**
Care		20	7	13	0.934068055	
C1						0.162281
	Disturbed	8	5	3	0.954434003	
	Not disturbed	12	2	10	0.650022422	
C2						**0.590176**
	Support	11	0	11	0	
	Does not support	9	7	2	0.764204507	
C3						0.0172375
	Firm	16	5	11	0.896038233	
	Lack of firm	4	2	2	1	

Table 5: Node Calculation Results 1.1.1

Node 1.1.1		**Amount**	**Decrease**	**Increase**	**Entrophy**	**Gain**
Caring-Not Support		9	7	2	0.764204507	
C1						**0.3197601**
	Disturbed	5	5	0	0	
	Not disturbed	4	2	2	1	
C3						0.0928885
	Firm	7	5	2	0.863120569	
	Lack of firm	2	2	0	0	

The following are the results of the calculations on the last node that obtained the decision to Increase in the Assertive attribute class and the decision to Decrease in the Less Assertive attribute class. Thus Node 1.1.1.1. is the last node formed.

Table 6: Node Calculation Results 1.1.1.1

Nodes 1.1.1.1		**Amount**	**Decrease**	**Increase**	**Entrophy**	**Gain**
Caring-Not Supporting-Not Disturbed		4	2	2	1	
C3						1
	Firm	2	0	2	0	
	Lack of firm	2	2	0	0	

2.6. Rapid Miner Test Results

The results of data processing with a decision tree model according to the Rapid Miner software, can be seen in Figure 3.1. as follows :

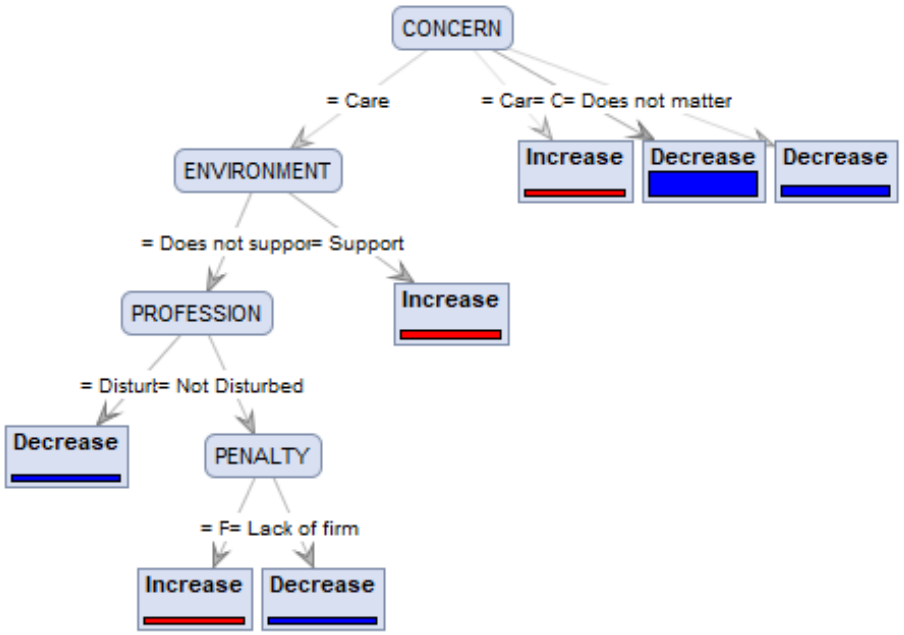

Figure 1: Result of decision tree

Furthermore, the final decision tree pattern is obtained after calculating and testing data on each attribute with the C4.5 Algorithm as follows:

Tree

```
CONCERN = Care
|   ENVIRONMENT = Does not support
|   |   PROFESSION = Disturbed: Decrease {Decrease=5, Increase=0}
|   |   PROFESSION = Not Disturbed
|   |   |   PENALTY = Firm: Increase {Decrease=0, Increase=2}
|   |   |   PENALTY = Lack of firm: Decrease {Decrease=2, Increase=0}
|   ENVIRONMENT = Support: Increase {Decrease=0, Increase=10}
CONCERN = Care : Increase {Decrease=0, Increase=1}
CONCERN = Careless: Decrease {Decrease=64, Increase=0}
CONCERN = Does not matter: Decrease {Decrease=16, Increase=0}
```

Figure 2: Description of decision tree

Based on data processing using RapidMiner software, the system accuracy value is 100%. Where the model that has been formed is tested for accuracy by using test data derived from training data with split validation on the rapidminer 5.3 application.

PerformanceVector

```
PerformanceVector:
accuracy: 100.00%
ConfusionMatrix:
True:   Decrease        Increase
Decrease:       26      0
Increase:       0       4
```

Figure 3: Value of Performance Vector Algorithm C4.5

3. Conclusion

Based on all the results of the research stages that have been carried out on the application of the C4.5 Algorithm in the analysis of the decline in public awareness in the application of the COVID-19 health protocol, it can be concluded that the problem of determining the dominant factor of decreasing public awareness can be solved using data mining techniques, namely the C4.5 Algorithm. Produce 6 rules and the level of accuracy generated by this method is 100% and from calculations with the C4.5 algorithm, the most dominant factor is Concern (C4) with a gain value of 0.3706246.

Reference

[1] J. Publicuho, D. Tuwu, and K. Kendari, “GOVERNMENT POLICY IN MANAGING THE COVID-19 PANDEMIC,” vol. 3, no. 2, pp. 267–278, 2020, doi:10.35817/jpu.v3i2.12535.

[2] EP Cynthia and E. Ismanto, "Decision Tree Algorithm Method C.45 in Classifying Sales Data," *J. Ris. Sis. inf. Dan Tech. information.*, vol. (3) July, no. July, pp. 1–13, 2018.

[3] Taufiq and Y. Yudihartanti, "Application of the C4.5 Algorithm Classification of Student Graduation Levels," *Semin. Nas. Computer Science.*, vol. 2, pp. 153–162, 2019.

[4] KF Irnanda, D. Hartama, and AP Windarto, “C4 Classification Analysis. 5 Against the Factors that Cause the Decline of Student Learning Achievement During the Pandemic Period,” vol. 5, pp. 327–331, 2021, doi:10.30865/mib.v5i1.2763.

[5] J. Jurnal and S. Information, “APPLICATION OF DATA MINING FOR WALLPAPER SALES PREDICTION USING C4 ALGORITHM. 5 STMIK Royal, Ksiaran,” vol. 2, 2016.

[6] IA Muis, M. Affandes, IA Muis, P. Studi, T. Informatika, and F. Sains, "Application of Support Vector Machine (SVM) Method Using Kernel Radial Basis Function (RBF) in Tweet Classification," vol. 12, no. 2, pp. 189–197, 2015.

[7] Z. Muhammad, "*Process Mining* School Academics Using Rapid Miner,” Vol. 2, no. 10, pp. 47–51, 2019.

[8] S. Case and S. Pringsewu, “ANALYSIS OF STUDENT SATISFACTION WITH ACADEMIC SERVICES USING THE C4 ALGORITHM METHOD. 5," vols. 02, no. 01, pp. 1–11, 2016.

[9] L. Farokhah, Y. Ubaidillah, and RA Yulianti, "Covid-19 Health Protocol Disciplinary Counseling in Gandul Village, Cinere District, Depok City," Semin. Nas. servant. Masy. LPPM UMJ, pp. 1–8, 2020.

[10] RK Sari, "Identification of the Causes of Citizens' Non-compliance with the Implementation of 3M Health Protocols During the Covid-19 Pandemic," J. AKRAB JUARA, vol. 6, no. 1, pp. 84–94, 2021.